

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets

(11)



EP 0 769 745 A1

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
23.04.1997 Bulletin 1997/17

(51) Int. Cl.⁶: G06F 11/34

(21) Application number: 96202620.9

(22) Date of filing: 19.09.1996

(84) Designated Contracting States:
DE FR GB IT SE

(72) Inventor: Senator, Steven T.
Colorado Springs, Colorado 80920 (US)

(30) Priority: 18.10.1995 US 544933

(74) Representative: Hanna, Peter William Derek et al
Tomkins & Co.,
5 Dartmouth Road
Dublin 6 (IE)

(71) Applicant: SUN MICROSYSTEMS INC.
Mountain View, California 94043-1100 (US)

(54) Device I/O monitoring mechanism for a computer operating system

(57) A device input/output ("I/O") monitoring mechanism is disclosed serving as an interface between a computer operating system kernel (26) and a device driver (30,32,34), which obviates the necessity of implementing specific pseudo-device drivers for various peripheral devices (36,38,40) and provides a standard interface between, for example, computer mass storage devices and a computer operating system. The I/O monitoring mechanism disclosed is of especial utility in the measurement of general storage device I/O performance and allows I/O statistics to be presented to application-level software (24) operating in conjunction with a computer operating system which facilitates the identification of storage device "hot spots" within meta-devices and sub-metadevices. The mechanism of the present invention is usable across a wide range of storage device peripheral hardware and device drivers and allows user-level (22) program code to readily determine the associated physical sub-component.

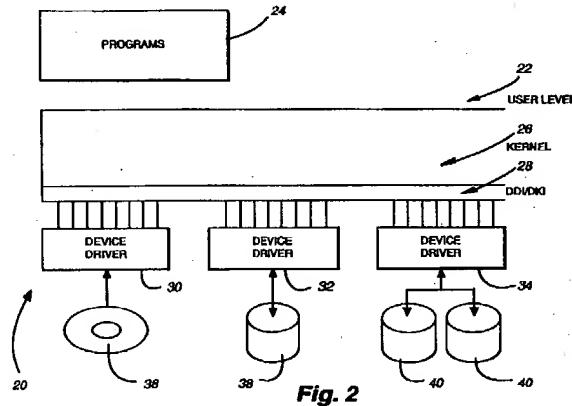


Fig. 2

EP 0 769 745 A1

Description**BACKGROUND OF THE INVENTION**

5 The present invention relates, in general, to the field of computer operating systems. More particularly, the present invention relates to a device input/output ("I/O") monitoring mechanism serving as a pseudo-device driver, or interface, between a computer operating system kernel and a device driver.

10 Device drivers comprise that portion of a computer operating system that contains hardware specific knowledge of a particular peripheral device. Such peripheral devices may include, for example, computer mass storage devices such as disk drives, mirrored storage devices, redundant arrays of inexpensive disks ("RAID"), hierarchical storage devices, virtual disks and the like. As a general rule, a particular device driver may, or may not, provide statistics back to the operating system kernel on device I/O operations according to its own predetermined metrics. Consequently, comparison of I/O statistics between devices may be rendered meaningless, very difficult or impossible due, respectively, to the possibly differing assumptions in the driver's metrics, specific implementations or actual unavailability of the data.

15 To this end, while pseudo-device drivers for interfacing between an operating system kernel and a device driver have previously been developed, *inter alia*, to monitor device I/O, they have been generally designed with a specific piece of hardware or software implementation in mind. That is, they have been specialized "piggy-back" drivers for coupling to a particular storage device or to traces generated from modified operating system software. As a result, specific knowledge of some portion of the hardware/software system being monitored was required in order to implement them.

SUMMARY OF THE INVENTION

20 The device I/O monitoring mechanism of the present invention obviates the necessity of implementing specific pseudo-device drivers for various peripheral devices and provides a standard interface between, for example, computer mass storage devices and a computer operating system. As a consequence, it is of especial utility in the measurement of general device I/O performance and can be utilized in conjunction with "opaque" (or "black box") devices without specific knowledge of their internal implementation.

25 More specifically, the device I/O monitoring mechanism of the present invention allows I/O statistics to be presented to application-level software operating in conjunction with a computer operating system which facilitates the identification of storage device "hot spots" within metadevices and sub-metadevices. The mechanism of the present invention is usable across a wide range of storage device peripheral hardware and device drivers and allows user-level program code to readily determine the associated physical sub-component.

30 In order to effectuate the foregoing, disclosed herein is a peripheral device I/O monitor for a computer having an operating system with a device statistics presentation mechanism resident thereon and also including a hardware abstraction layer for interfacing to an associated peripheral device. The peripheral device includes a device driver operatively responsive to the hardware abstraction layer of the computer operating system. The device I/O monitor comprises a pseudo-device driver operatively interposed between the hardware abstraction layer of the computer operating system and the device driver and itself presents corresponding first and second pseudo-device driver hardware abstraction layers. The first pseudo-device driver hardware abstraction layer is coupled to the computer operating system and the second pseudo-device driver hardware abstraction layer is coupled to the device driver.

35 In accordance with a method of the present invention, I/O metrics for a peripheral device coupled to a computer may be determined wherein the computer includes a computer operating system incorporating a device driver for the peripheral device. The method comprises the steps of: intercepting one or more selected calls from the computer operating system to the device driver; providing for initiating a call back to a portion of the operating system to record an occurrence of the one or more selected calls; and providing for passing the one or more selected calls to the device driver. In accordance with a more specific method, the method may comprise the additional steps of: further intercepting one or more selected device calls from the device driver to the computer operating system; further initiating an additional call back to the operating system to record an occurrence of the one or more selected device calls; and further passing the one or more device calls to the operating system. The one or more selected calls may advantageously be "read", "write" or "strategy" operations when utilized in conjunction with a UNIX® System V compliant operating system.

BRIEF DESCRIPTION OF THE DRAWINGS

40 The aforementioned and other features and objects of the present invention and the manner of attaining them will become more apparent and the invention itself will be best understood by reference to the following description of a preferred embodiment taken in conjunction with the accompanying drawings, wherein:

45 Fig. 1 is a simplified representational drawing of a general purpose computer forming a portion of the operating environment of the present invention;

Fig. 2 is a simplified representational drawing of a computer operating system illustrating the kernel and various device drivers as interfaced by a Device Driver Interface/Device Kernel Interface ("DDI/DKI") hardware abstraction layer in a conventional configuration and further illustrating various types of computer mass storage devices and their associated drivers;

5 Fig. 3 is an additional simplified representational view of the computer operating system and driver of Fig. 2 and further illustrating the kernel statistics ("KSTAT") module within the kernel of a particular operating system which functions as a device driver statistics module for the peripheral device drivers;

10 Fig. 4 is a further simplified representational view of a computer operating system including a DDI/DKI interface and a device I/O monitoring mechanism in the form of a pseudo-device driver interposed between the operating system and a device driver for updating the KSTAT module in the event of the occurrence of certain I/O access operations between the operating system and a device driver; and

15 Fig. 5 is a more detailed illustration of the sequence of calls and call-backs initiated between the operating system and device driver through the I/O monitoring mechanism of Fig. 4 during a representative access to a computer mass storage device during operation thereof.

DESCRIPTION OF A PREFERRED EMBODIMENT

The environment in which the present invention is used encompasses the general distributed computing system, wherein general purpose computers, workstations or personal computers are connected via communication links of various types, in a client-server arrangement, wherein programs and data, many in the form of objects, are made available by various members of the system for execution and access by other members of the system. Some of the elements of a general purpose workstation computer are shown in Fig. 1, wherein a processor 1 is shown, having an input/output ("I/O") section 2, a central processing unit ("CPU") 3 and a memory section 4. The I/O section 2 is connected to a keyboard 5, a display unit 6, a disk storage unit 9 and a compact disk read only memory ("CDROM") drive unit 7. The CDROM unit 7 can read a CDROM medium 8 which typically contains programs 10 and data. The computer program products containing mechanisms to effectuate the apparatus and methods of the present invention may reside in the memory section 4, or on a disk storage unit 9 or on the CDROM 8 of such a system.

With reference now to Fig. 2, a simplified representational view of a computer operating system 20 is shown. The computer operating system 20 includes a user level 22, at which level various application programs 24 may be run. The user level 22 overlies a kernel 26 which includes, in the case of a UNIX System V compliant operating system such as Solaris™, developed and licensed by Sun Microsystems, Inc. a DDI/DKI interface 28 as will be more fully described hereinafter. The DDI/DKI interface 28 couples the kernel 26 to a number of device drivers, such as the representative device drivers 30-34 which contain hardware specific code regarding particular peripheral devices such as the computer mass storage devices illustrated, wherein device 36 represents a CDROM, device number 38 represents an industry standard disk drive and devices 40 represent a pair of mirrored disk drives.

The illustration of Fig. 2 depicts a generalized System V compliant operating system 20 incorporating a DDI/DKI interface 28 hardware abstraction layer ("HAL") or hardware interface layer. The device drivers 30-34 incorporate the specialized code necessary to communicate with the corresponding devices 36-40. Utilization of a standard hardware abstraction layer such as the DDI/DKI interface 28 allows the operating system to be most readily interfaced to drivers 30-34 specifically written for communication with respective devices 36-40 of differing manufacturers and the interface between the drivers and the kernel 26 remains the same regardless of the device characteristics. In System V compliant operating systems, peripheral device "reads", "writes" and other special operations are denominated as I/O control ("ioctl") and any specific operation for a given device 36-40 (as to its sectors, platters, read/write heads and the like) is coded as a part of the corresponding device driver 30-34 in order to enable the operating system 20 to select the data requested. Generally, computer mass storage devices deal in chunks of data called "blocks". However, the application programs 24 relate data to named files or persistent objects and the blocks do not have specific file names associated with them. Calls to "open" and "read" a particular file are routed through the operating system 20 file system which then maps the file names to particular blocks on the storage device.

With reference additionally now to Fig. 3, the interaction between a generic device driver 42 and the KSTAT module 46 of the operating system 20 kernel 26 is shown. A "black box" driver 40 having unknown characteristics for interface to a specific device (not shown) is also illustrated. Structure and layers analogous to that previously described with respect to the preceding figures is like numbered and the foregoing description thereof shall suffice herefor.

As previously described, the layer between the kernel 26 and the device drivers 42, 44 is called the DDI/DKI interface 28. The DDI/DKI interface 28 is strictly defined, inclusive of specific routines that call into a portion of the operating system 20 called the kernel statistics ("KSTAT") module 46. Included in the DDI/DKI interface 28 are a number of interfaces, or methods, to call in from a device driver 42, 44 into that portion of the kernel 26. Among the things that a device driver 42, 44 may communicate to the kernel 26 is how many I/O operations it has outstanding, how long it's taking for those operations to be completed, the size of those operations, and the types (i.e. "read" or "write") of those operations. By supplying these generic interfaces to the KSTAT module 46, the device drivers 42, 44 do not have to undertake I/O

bookkeeping themselves. Utilizing the KSTAT module 46, the device drivers 42, 44 merely have to communicate a call and set appropriate counters in the KSTAT module 46 which then addresses the presentation of the I/O data to the user.

In the case of a "black box" device driver 44 associated with an unknown computer mass storage device, it may only be known that it somehow provides data storage and plugs into the DDI/DKI interface 28. However, nothing may be known about its I/O operations. In that case, it cannot be determined whether or not the unknown device driver 44 is any better or any worse than the device driver 42 and it has heretofore been impossible to determine how many I/O operations are outstanding, how long they are taking and the like rendering it very difficult to characterize the unknown device driver 44.

With reference additionally now to Fig. 4, the various calls which may be made through a standard DDI/DKI interface 28 associated with the operating system 20 kernel 26 are shown in conjunction with the pseudo-device driver 50 of the present invention. As an example, an operating system call 60 is issued which is intercepted by the pseudo-device driver 50 to make a callback 62 to the KSTAT module 46. The call 60 then proceeds through as a passed-through call 64 to the device driver 42. Likewise, a device call 66 is intercepted from the device driver 42 and, in response, a call to the KSTAT module 46 is issued by the pseudo-device driver at 68. The device call 66 is then passed through as a passed through call 70.

In addition to the "ioctl" call mentioned previously, the kernel 26 may issue a call to "open" which instructs the device driver 42 to open, for example, partitions. On the other hand, the "close" instruction indicates that whatever operation has been undertaken since the "open" (i.e. "read" where the mass storage device sectors actually begin) has been completed. Alternatively, the device driver 42 may function to pre-position the data transducers of an associated mass storage device to anticipate a subsequent "read" or "write" command. As illustrated, System V compliant operating systems also include a "strategy" routine which is a type of generalized "read" or "write" operation. The designations "aread" and "awrite" designate asynchronous reads and writes respectively.

The pseudo-device driver 50 of the present invention provides hooks into the DDI/DKI interface 28 and, itself presents first and second DDI/DKI interfaces between the operating system 20 and the device driver 42. These hooks allow it to provide information back to the KSTAT module as to how many I/O operations are currently in progress, how long they have been taking and the length of the queues, etc.

In the example shown, if the kernel 26 issues an operating system call 60 constituting a "read" to the device driver 42 associated with a particular computer mass storage device or subsystem, the pseudo-device driver 50 recognizes that event and makes a callback 62 to the KSTAT module 46 to increment a counter to record that the read call 60 has been issued. The pseudo-device driver 50 then passes the call 60 on to the device driver 42 as shown by the dashed line 64. When the "read" operation is completed, the device driver 42 issues a device call 66 which is detected by the pseudo-device driver 50 to then issue an additional call to the KSTAT module 46 signifying the completion of this particular I/O operation. As before, the pseudo-device driver 50 then passes the device call 66 through to the kernel 26 as indicated by dashed line 70 completing the operation.

In this manner, "black box" device drivers (such as device driver 44 of Fig. 3) can be measured such that it may be accurately compared with competitive drivers. Moreover, since the pseudo-device-driver 50 identifies and monitors I/O operations of particular interest, it can provide information to the operating system 20 KSTAT module 46, and hence to application programs regarding I/O metrics. In fact, more than merely generalized information about a particular peripheral may be monitored and a given disk drive or disk subsystem may be treated as a succession of smaller disk drives to provide statistics on particular cylinders, regions, platters or the like to determine "hot spots" or where the device is fastest or, conversely, where it is slowest. This information may, of course, then allow I/O operations to be selectively moved around to optimize overall performance or merely provide confirmation that a device may not be operating properly (such as, for example, the identification of a number of "re-try's on particular blocks). Stated another way, the pseudo-device driver 50 is an enabler for other device monitoring tools through the existing KSTAT module 46 while remaining essentially transparent to calls between the operating system 20 and a given device driver 42. This is of especial utility in conjunction with databases which would desirably be able to have their index tables located at the fastest portion of a particular storage device or subsystem to speed operations attempting to access data pointed to by the table. Alternatively, the pseudo-device driver 50 and the detailed knowledge it can accumulate regarding I/O operations and specific device functionality would enable the relocation of large graphics files to storage areas that exhibit greater speed for large I/O operations.

The pseudo-device driver 50 allows information to be supplied to the KSTAT module 46 about other than just traditional "open" and "close" calls but also, for example, with respect to "reads" "writes" and "strategy" calls without specific knowledge of the particular device driver 42 then being accessed or whenever a specific command is passed through it and about which it has no other knowledge. The pseudo-device driver 50 is hot designed to "error off" as a normal device driver would upon receipt of an unrecognized command and such a command would be, instead, merely passed on to the device driver 42 for it to return an error back if appropriate, which error is then passed back up through to the kernel. "Done" routines may also be captured and recognized by the pseudo-device driver 50. Upon receipt of an I/O call of interest, the pseudo-device driver 50 increments the appropriate counters in the KSTAT module 46 and then takes note of the times when they are occurring and their "done" times in order to calculate the duration of that particular

operation.

As an example, private I/O control may be provided to record I/O information about a particular disk drive, or a region of a particular drive. In operation, the pseudo-device driver 50 is set up (or administered) as a device driver and the type of information to be monitored is then defined. As part of the administration of the pseudo-device driver 50 of the present invention, it may be loaded into the kernel 26 as with conventional device drivers. The operations to be monitored are then defined with respect to the actual device, the number of regions or size of the regions to be monitored and the like.

With reference now to Fig. 5, a further exemplary scenario is illustrated in conjunction with the pseudo-device driver 50 of Fig. 4 to more fully illustrate the operating system level and device driver level DDI/DKI interfaces 80, 82 thereof.

At line (a) a "read" operation is initiated, with the assumption that "reads" are operations of interest to the pseudo-device driver 50 as administered. Operations which are not of interest are passed directly through the pseudo-device driver 50, without intervention, to the device driver. Since the pseudo-device driver 50 has been administered to react to "reads", at (b) the KSTAT module is incremented and the time recorded directly or by I/O operation number in order to later reconstruct the time of its initiation.

At (c), the "read" operation is passed on down to the device driver which then causes a "read" operation to occur in a conventional manner. Upon completion of the operation at (d), there is a call-back made and it presents the data read from the mass storage device.

At (e), an end of operation signal is sent to the KSTAT module which is referenced either to time or operation identifier and/or in conjunction with a decrementing of the number of current operations outstanding and an incrementing of the count of operations completed. In operation, the pseudo-device driver 50 is transparent to the intended interaction between the kernel and the device driver with only confirmation to the KSTAT module when something of interest is received from the kernel that's intended for the driver. The pseudo-device driver also informs the KSTAT module when something is received from the driver that is intended for the kernel. The same scenario holds true for "writes" in addition to "reads" with only the direction of the data transfer being different.

"Strategy" operations are similar to "read" and "write" operations with the difference primarily being that fixed size I/Os are generally encountered. At (f) a call is directed back to the device driver which then at (g) initiates the actual "done" message.

In general, "read", "write" and "strategy" I/O operations will be of interest in determining device metrics. On the other hand, I/O "open", "close", "ioctl", "aread" and "awrite" may not be considered particularly interesting (although their measurement may be of interest in other applications) and are merely passed through to the device driver.

With reference to the following Table 1, CB_OPS_VECTOR operations are shown which map the new "read", "write" and "strategy" operations in conjunction with the pseudo-device driver 50 of the present invention. Other operations remain unchanged. The table illustrates entries which are labeled "open", "close" etc. with their values corresponding to the names of the device upon which the operation is to occur (i.e. "disk open"). With respect to the "read", "write" and "strategy" operations, the actual routines to be called are shown. The values of these routines include the address in memory to which to direct the program control flow.

TABLE 1

OPEN	CLOSE	READ	WRITE	STRATEGY	IOCTL	AREAD	AWRITE
disk_OPEN()	disk_CLOSE()	NEW_READ()	NEW_WRITE()	NEW_STRAT()	disk_IOCTL()	disk_AREAD()	disk_AWRITE()

45

In the specific embodiment described above, the Solaris® operating system, available from Sun Microsystems, Inc.,

DDI/DKI interface 28 defines the KSTAT module 46 interface. This provides a mechanism for I/O statistics to be presented to application-level code in a standard format which is defined in the kstat(3K) manual page under the description of KSTAT_IO_TYPE kernel statistics.

In order to present this data, calls into the KSTAT module need to be made in driver "strategy" and "done" routines. Because the buffer header b_iodone field is not a chain but a replacement done routine, it is necessary to wrap transactions in a separate buffer header. The strategy routine is redirected to the pseudo-device driver 50 strategy routine to increment appropriate kstat counters in the KSTAT module 46, but otherwise it is intended to be a "pass-through" operation. All other entry points simply pass through to the underlying device 44.

Intra-device statistics are most useful when they can be used to identify hot spots. This requires a mechanism to define the granularity of the chunks on which statistics will be taken. The pseudo-device driver 50 includes an "ioctl" which defines the size of these extents. Limiting factors on this are the number of bits in a daddr_t. This allows the measurement to be made down to the block level, although the defaults may be much larger, depending upon the appli-

cation accessing the device. When the extent sizes are defined, kstat_io_type structures may be created and initialized. Presentation of this data can be done by "metatool", "iostat" or another mechanism and need only comply with the current kstat/kvm library interface.

I/O history, and thus determination of device "hot spots", is ultimately the responsibility of user-level code with all of the implementation and customizability that implies. One significant advantage of the present invention is that it allows this mechanism to be used on top of "opaque" or "black box" devices. In general, a mutex may be required around all pseudo-device driver 50 operations in order to prevent module unloading while active. However, the impact of this can be minimized by using multiple mutexes and counters incurring the overhead upon attempts to modunload, rather than on every call into this layer.

While there have been described above the principles of the present invention in conjunction with specific operating systems, device drivers and peripheral devices, the foregoing description is made only by way of example and not as a limitation to the scope of the invention. Particularly, it should be noted that the principles of the present invention are applicable to the IBM AIX™ and MicroSoft Windows NT™ operating systems (or other UNIX® System V compliant operating systems in addition to the Sun Microsystems, Inc. Solaris™ implementation shown) as well as any other operating system that provides a generic kernel to device driver interface in conjunction with device statistics presentation mechanism.

Claims

1. A method for determining input/output (I/O) metrics for a peripheral device (44) coupled to a computer, said computer including a computer operating system (20) incorporating a device driver (42) for said peripheral device, said method comprising the steps of:

providing for intercepting one or more selected calls (60) from said computer operating system to said device driver;

providing for initiating a call back (62) to a portion (46) of said operating system to record an occurrence of said one or more selected calls; and

providing for passing (64) said one or more selected calls to said device driver.

2. The method of claim 1 wherein said one or more selected calls (60) is an access to said peripheral device selected from a group comprising read, write and strategy operations.

3. The method of claim 1 further comprising the steps of:

providing for further intercepting one or more selected device calls (66) from said device driver (42) to said computer operating system (20);

providing for further initiating an additional call back (68) to said portion (46) of said operating system to record an occurrence of said one or more selected device calls; and

providing for further passing (70) said one or more device calls to said operating system.

4. A peripheral device I/O monitor for a computer having an operating system (20) resident thereon with a device statistics presentation mechanism and including a hardware abstraction layer (28) for interfacing to an associated peripheral device (36,38,40,44) and a device statistics presentation mechanism (46), said peripheral device including a device driver (30,32,34,42) operatively responsive to said hardware abstraction layer of said computer operating system, characterised by:

a pseudo-device driver (50) operatively interposed between said hardware abstraction layer of said computer operating system and said device driver, said pseudo-device driver presenting corresponding first and second pseudo-device driver hardware abstraction layers (28), said first pseudo-device driver hardware abstraction layer being coupled to said computer operating system and said second pseudo-device driver hardware abstraction layer being coupled to said device driver.

5. The device I/O monitor of claim 4 wherein said computer operating system (20) is a UNIX System V compliant operating system.

6. The device I/O monitor of claim 4 wherein said first and second pseudo-device driver hardware abstraction layers (28) are Device Driver Interface/ Device Kernel Interface (DDI/DKI) layers.

7. The device I/O monitor of claim 4 wherein said pseudo-device driver functions to intercept one or more selected

calls (60) from said computer operating system (20) to said device driver (42) and initiate a call back (62) to said device statistics presentation mechanism (46) through said computer operating system and said first pseudo-device driver hardware abstraction layers to record an occurrence of said one or more selected calls.

- 5 8. The device I/O monitor of claim 7 wherein said pseudo-device driver further functions to pass (64) said one or more selected calls to said device driver (42) through said second pseudo-device driver hardware abstraction layer.
- 10 9. The device I/O monitor of claim 8 wherein said one or more selected calls comprise an access to said peripheral device selected from a group comprising read, write and strategy operations.
- 15 10. The device I/O monitor of claim 7 wherein said pseudo-device driver functions to further intercept one or more selected device calls (66) from said device driver (42) through said second pseudo-device driver hardware abstraction layer to said computer operating system (20) and further initiate an additional call back (68) to said device statistics presentation mechanism (46) through said computer operating system and said first pseudo-device driver hardware abstraction layers to record an occurrence of said one or more selected device calls.
- 20 11. The device I/O monitor of claim 10 wherein said pseudo-device driver functions to further pass (70) said one or more device calls (66) to said operating system through said computer operating system and said first pseudo-device driver hardware abstraction layers.

20

25

30

35

40

45

50

55

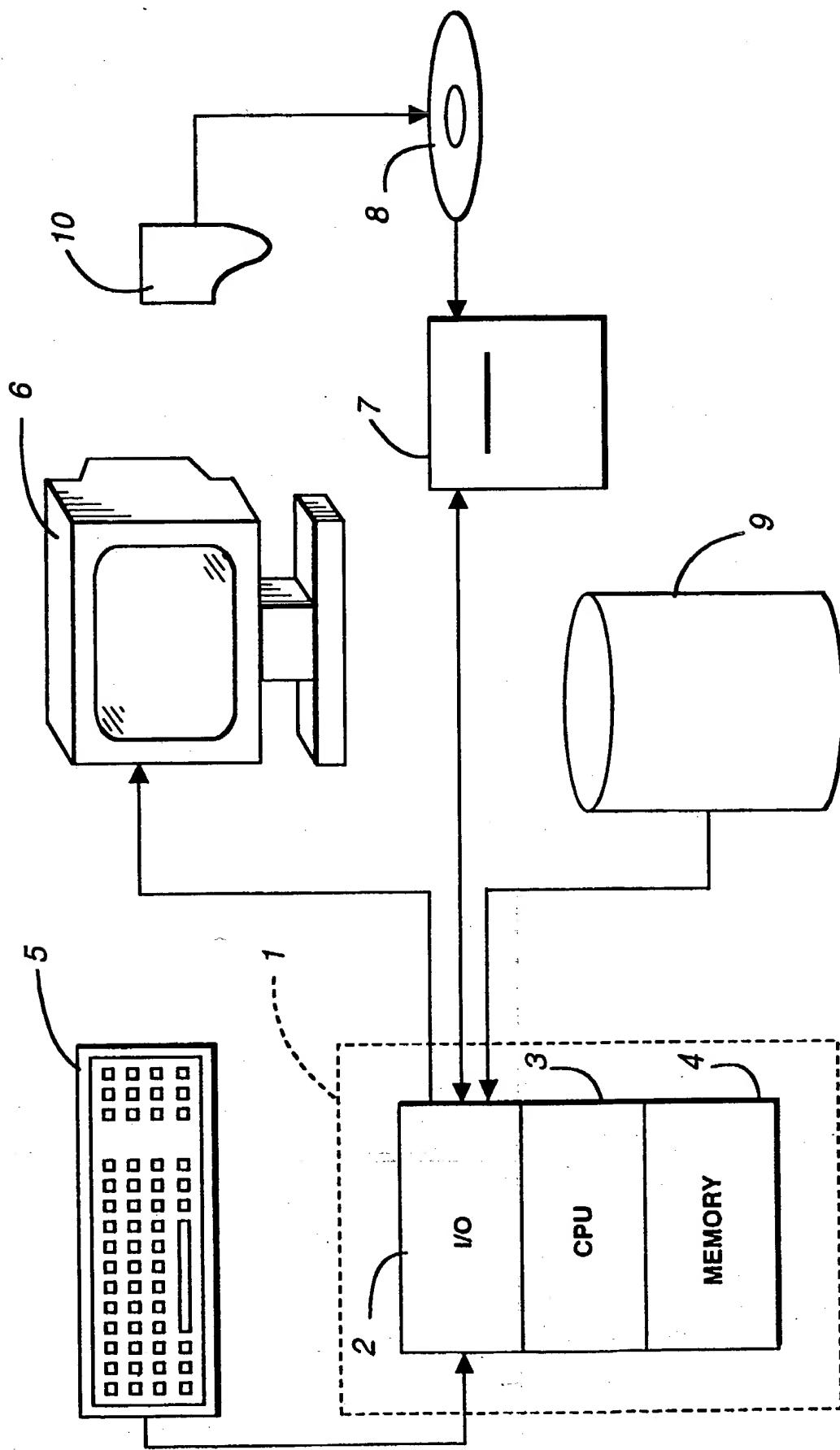
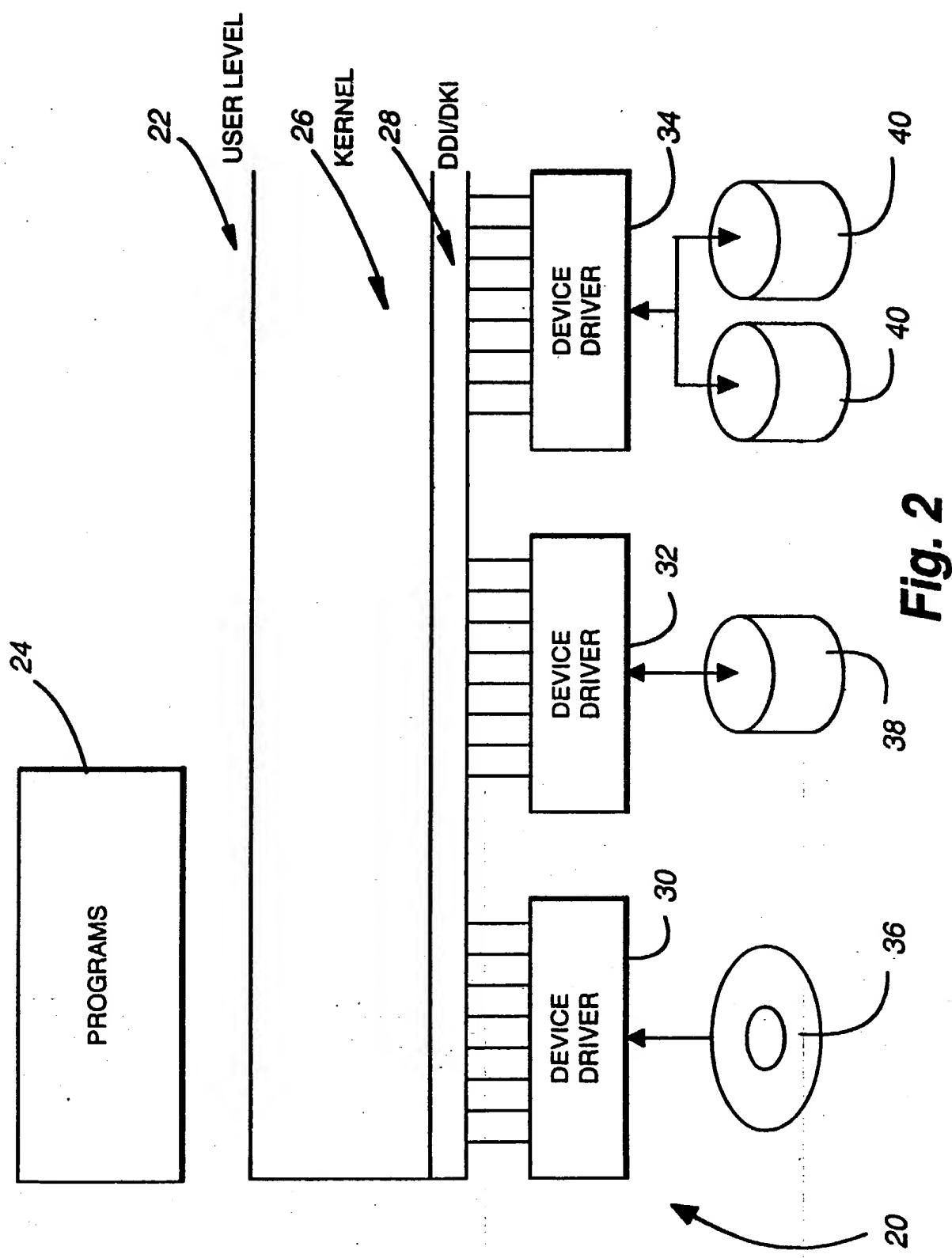


Fig. 1

**Fig. 2**

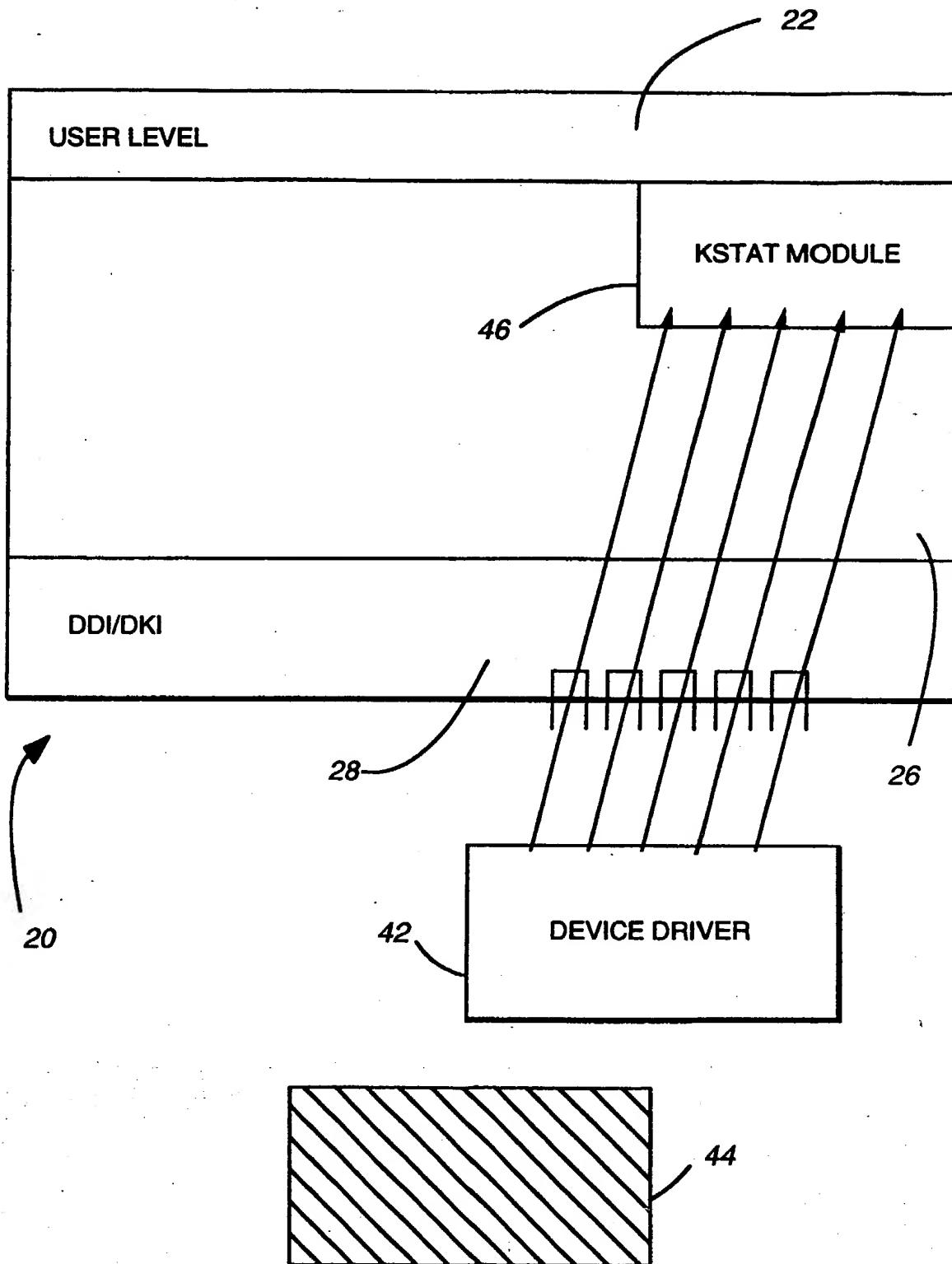
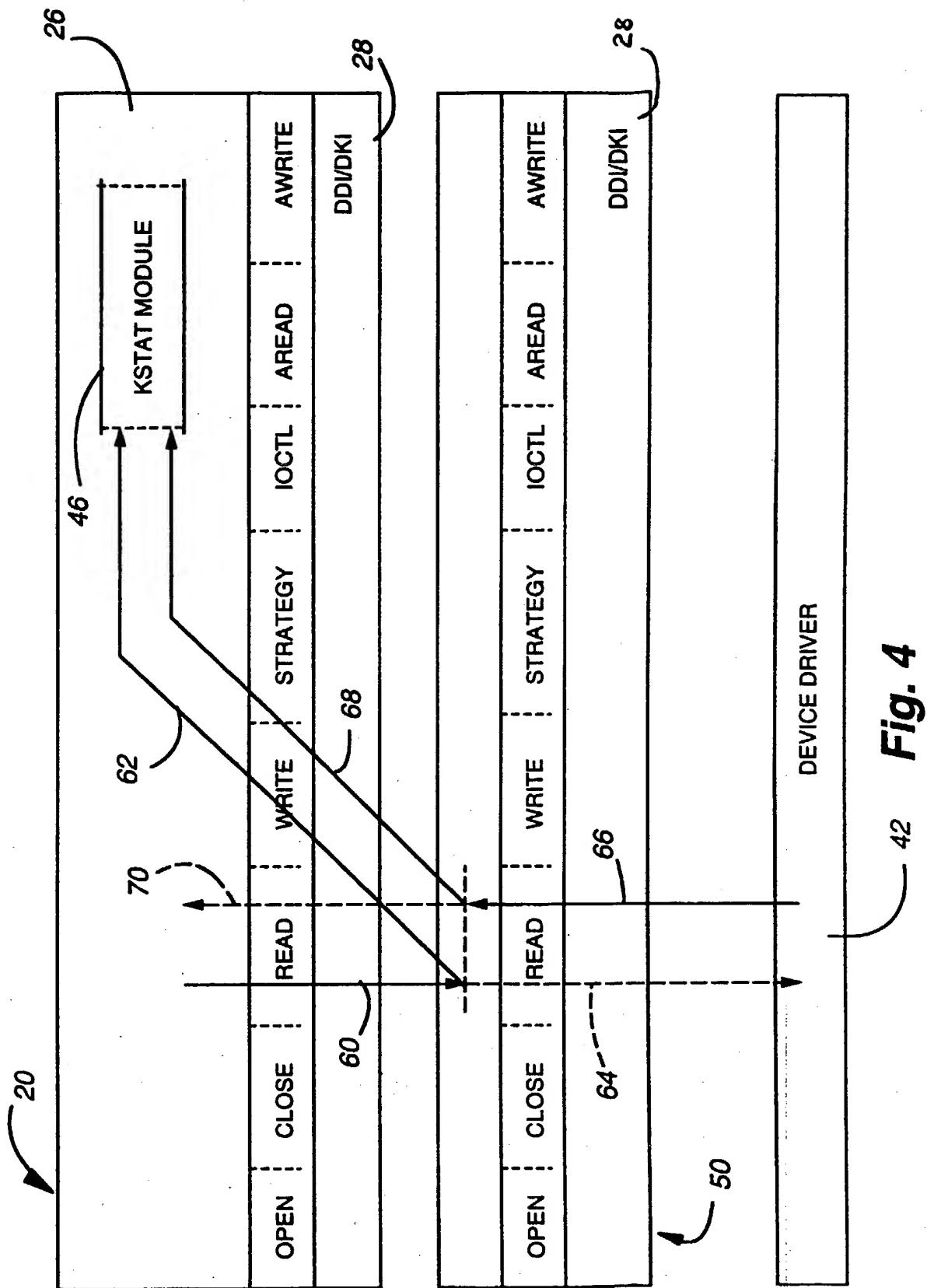


Fig. 3

**Fig. 4**

42